



Corporate Online/ CD-ROM Publishing: The Design and Tactical Issues

Typically, corporate publishing efforts involve many different kinds of documents, which may change frequently over the course of a corporation's long-term investment in the content being retrieved and re-cast.

Existing document imaging software effectively serves most small business needs when it comes to publishing documents from multiple sources in various formats on CD-ROM and the World Wide Web, even when such projects are undertaken on shoestring budgets. Large-scale corporate publishing, however, presents a different set of problems.

Typically, corporate publishing efforts involve many different kinds of documents, which may change frequently over the course of a corporation's long-term investment in the content being retrieved and re-cast. Requirements of corporate in-house document publishing typically include the following:

- Avoiding hand-crafting documents for different media;
- Having the flexibility not to be hostage to changing word processors, vendor alliances, operating systems, or output media;
- Reducing exception handling as volumes of published documents increase;
- Incorporating—if Web publishing is planned—support for upcoming changes in HTML.

Rules are easy enough to state theoretically, but often much harder to live by. The first step in this level of "corporate good living" is to analyze the state of the documents and the publishing culture of the organization developing the electronic publishing program.

Second, if it is not already done, the company needs to discipline and manage a single-source document base, filtering from it various documents as needed. The third step is careful analysis of the organization's electronic document needs, which will doubtless mean re-engineering any existing publishing processes and data. Comparing the major choices and tradeoffs, based on a fundamental understanding of the key technologies and concepts, may be the best way to make the most of the document-to-delivery data imaging and retrieval phase of corporate online and CD-ROM publishing.

SGML AND HTML: DOCUMENT DEFINERS

Although Standard Generalized Markup Language (SGML) has been an ISO standard since 1986, corporate publishers in some industries are still not sure what the acronym stands for, or how to use it. Describing SGML is like reciting a myth-dispelling litany: SGML is not a graphical standard. You can't "buy SGML." SGML is not—in spite of its name—even a markup language, *per se*; rather, it is a high-level set of rules for creating markup languages. SGML is an ISO standard defined at such a sophisticated level that it can be used to describe many things besides documents, and is the basis for newer electronic publishing and multimedia standards (such as HyTime, ISO 10744).

Markup languages are expressed, following SGML rules, in "Document Type Definitions" (DTDs). SGML's original intent was to describe only document

content, such as section headings and table cells, and not document *format*. DSSSL (Document Style Semantics & Specification Language, ISO 10179) provides a standard architecture for formatting specification. However, SGML is such a flexible standard that it can be used to describe whatever is needed. Consequently, implementing SGML can prove tremendously useful for corporate institutions and others interested in developing in-house electronic publishing programs, in particular because it enables businesses to do the following:

- Compress time to market;
- Reduce life cycle costs;
- Produce multiple versions in multiple media of documents from a single document base;
- Gain some independence from vendors, changing file formats, operating systems, and the like;
- Create document content which lasts as long as needed.

A document's DTD, in short, is its rosetta stone. With a document and its DTD, you, as a would-be corporate electronic publisher, can decipher the document's content unambiguously, and store the information it contains in only one place and in only one form. By using a sufficiently detailed DTD to describe your documents, you can create them once and filter out many different versions on a variety of media.

HyperText Markup Language (HTML) is a particular brand of SGML document type definition. Designed in 1990 and first implemented in 1992, HTML is now used more frequently as a display language—concerned primarily with the way information looks on a screen—than a language expressing document structure. Although HTML originally described some structures, they were usually simple ones like paragraphs and headings. The emphasis was on document appearance displayed on Web browsers, with particular attention to enhancements like boldface and italic.

Although it will retain older tags so documents using any vintage of HTML can still be displayed on the Web, HTML is upgrading its content structures dramatically. The language's third incarnation also provides some very robust content modeling of document

information, including tables, math, equations, and improved forms.

CORPORATE PUBLISHING STRATEGIES: DATA AND PROCESSES

SGML usually deals with content, not form, so even when a DTD defines a visual element, such as a horizontal rule in HTML, it treats it as a well-defined object. Printer description languages help supplement document imaging needs where SGML falls short, and two such languages survive in the market today: Adobe's PostScript and HP's PCL. For all its virtues, however, PostScript has several shortcomings. For one, showing the result of a PostScript—"Display PostScript"—was never feasible on a large scale. Interpreting PostScript and displaying it onscreen required too much CPU power. Moreover, PostScript provides no hypertext linking, contains no font metrics, and is linear—i.e., to use it you must start at the beginning and go to the end.

Adobe's "second-generation PostScript," called Portable Document Format (PDF), displays printed output precisely onscreen. And PDF files, like PostScript files, are cross-platform. Moreover, because Adobe has made PDF open, many vendors are developing plug-in software to work with it, and Adobe itself is developing or distributing plugins for its own products.

But first, a reality check: 80 percent of legacy documents come in the two forms of paper and typesetting files, which means that most corporations standardize on publishing systems and word processors, but do not expend much energy assuring the uniformity of the documents at the binary level. Everyone is pressed for time, and most learn only enough about any one tool to produce attractive paper results.

If your corporation is in this category, divide and conquer your publishing data. For documents like reference or service manuals that can benefit from SGML, develop an applicable DTD. To preserve your Web publishing options, design your DTD so that it can work with HTML Version Three. Then get that data re-keyed or converted to SGML-tagged files.

Vendors are available who will learn your DTD and apply it to your legacy data for \$1.20 to \$1.50 per 1,000 characters.

The trade-off you will have to weigh is richness of DTD—capturing structure in all its detail—versus the increasing percentage of re-keying of what will be SGML tags, not data itself.

Using native SGML authoring tools produces valid SGML, no matter how complex the DTD, so corporations committed to SGML often decide to use these specialized tools. However, the costs of such tools, and the additional costs of the retraining and support their implementation will require are sometimes more than corporations want to pay. These companies already are comfortable with one or more word processors, and the prevailing trend finds word processors becoming increasingly SGML-aware. MS Word, for instance, now offers Internet Assistant for converting native Word files to HTML. Additionally, Microsoft's SGML Author for Word lets authors tag and validate SGML documents using Word for Windows styles. WordPerfect has developed a powerful DOS and UNIX SGML product called Intellitag that should debut with version 6.1, SGML Edition. SGML awareness should continue to increase in word processors, as the desire for Web publishing persists unabated.

However, be wary. If authors do not use these upscale word processors correctly, SGML may prove of little use. For example, if authors do not understand the difference between true tables and tabbed columns, SGML derived from these documents will not take advantage of upcoming HTML capabilities.

WEB/CD-ROM PUBLISHING PRODUCTS: WHAT'S AVAILABLE NOW

Having developed documents in SGML, you'll be able to re-purpose their content on CD-ROM or the net with a growing number of SGML tools. An early industry leader was Electronic Book Technologies (EBT) in Providence, Rhode Island, whose first product, DynaText, was a viewer capable of displaying any SGML DTD, much like a super Web browser. EBT's customers routinely publish on CD-ROMs, LANs, and the Web. The upfront investment in products that use SGML pays handsome dividends, such as automatic hyperlinks created for SGML tags, and the ultimate in flexible re-purposing of content.

EBT's Web-specific companion piece to DynaText, DynaWeb, gives Web publishers the means to distribute information products using the same tools and content as for CD-ROM and LAN publishing. The DynaWeb server makes existing DynaText books available to Web clients with essentially no extra effort. Furthermore, DynaWeb gives users access to very large SGML documents—hundreds of times larger than HTML files—and supplies sophisticated DynaText SGML searching via HTML+ query forms. And since EBT products were designed from the beginning to be able to handle any SGML documents, they should be compatible with any future version of HTML.

Hot on the heels of EBT's products are many others. One impressive option is SoftQuad's Panorama Pro, a general-purpose SGML publishing tool, that emphasizes flexible authoring and use of Web documents. With a growing demand for multimedia documents, Panorama Pro is HyTime-ready, which means it can exploit the power of SGML for multimedia.

If your firm has a publishing system using SGML files but also insists on rich layout for certain documents, you can produce PDF for these files. Such systems allow formatting information to be associated with each SGML tag. Simply output to PostScript and then to PDF, or to PDF directly.

Upcoming versions of Netscape will let users view large PDF files in page-size chunks, rather than forcing them to wait for multimegabyte files to download. Products such as InContext have begun to develop hybrid tools which combine the power of SGML and Acrobat. And as Web browsers become general-purpose SGML browsers (suitable for use with documents on the Web, CD-ROM, or LANs), HTML itself may become merely the lowest common denominator among Web languages.

SGML AND HTML-BASED PUBLISHING: EVOLUTION OR REVOLUTION?

As Web publishers know, 5MB of document data might easily become 100 separate HTML files. Hypertext, electronic

media, and limited network bandwidth cry out for modularity in documents and better linking between them. HyTime, another SGML standard, promotes enhanced linking across documents and media types on any media.

Furthermore, authoring practices must change far more than the authors simply mastering new tools. But before you, as a new initiate in corporate electronic publishing, go that next step, making significant investments in more advanced tools and learning and training staff in new authoring approaches, make sure you understand the tradeoffs between tactical and strategic decisions in the emerging corporate publishing world.

Robert J. Boeri is Advanced Systems Specialist in the Information Services Division of Factory Mutual Engineering, Norwood, MA. He may be contacted via Internet—bboeri@world.std.com. Martin Hensel is president of Martin Hensel Corporation, a consulting firm that builds SGML-based editorial and production systems for publishers, corporations, interactive services, and composers. Hensel founded LaserData in 1981 and was chairman of the NISO committee that developed ISO 9660. He may be contacted at Martin Hensel Corporation, 233 Needham Street, Newton, MA 02164; 617/527-3230; Fax 617/527-1929; Internet—martin.hensel@hensel.com. ■